Домашние задания курса по Трансформерам (Сбер)

Введение

- Пожалуйста, прочтите документ очень внимательно. Если у вас есть вопросы, задавайте их в группе telegram.
- Используйте данный шаблон для оформления обоих заданий: https://colab.research.google.com/drive/1ttPT6X4K0ovgbzmNjlcEiprkj1LaBuF2#s=
- Для обеих задач вам необходимо отправить свои решения в Codalab (ссылки можно найти ниже).
- Необходимо использовать "**SberSummer2023**" в качестве названия вашей команды

Semantic role labelling: https://codalab.lisn.upsaclay.fr/competitions/531
Detoxification: https://codalab.lisn.upsaclay.fr/competitions/642

Задание 1 - Semantic Role Labelling

1.1 Введение

Задача мотивирована потребностями человека сравнивать различные объекты: разные модели мобильных телефонов, автомобилей, языки программирования, страны и т.д. Исследования NLP частично решают данную задачу сравнения объектов, однако в настоящее время существует множество возможностей для улучшения существующих сравнительных систем ответов на вопросы.

Например, система CAM (Comparative Argument Mining) получает пару объектов для сравнения и извлекает аргументы в пользу каждого из них. Аргументами в сравнительных предложениях служат предикаты (сравнительные характеристики объектов, например, *проще, лучше, быстрее* и т.д.) и аспекты (характеристики, по которым сравниваются объекты, например, *скорость, экран, производительность* и т.д.). Аспекты и предикаты извлекаются с использованием рукописных шаблонов, которые имеют низкую полноту (Recall) – не удается извлечь объекты, которые не

соответствуют шаблонам, – а иногда извлекаются неправильные объекты. В данной задаче вам предлагается улучшить процесс извлечения аргументов (объектов, аспектов и предикатов) из предложения. Такая модель должна быть обучена на предложениях, где слова или фразы имеют разметку последовательности – каждому слову соответствует его тег.

Примеры предложений

> Postgres is easier to install and maintain than Oracle.
>
> *[Postgres OBJECT] is [easier PREDICATE] to [install ASPECT] and [maintain ASPECT] than [Oracle OBJECT].*

Сущности могут состоять из нескольких слов:

> Advil works better for body aches and pains than Motrin.
>
> *[Advil OBJECT] works [better PREDICATE] for [body aches ASPECT] and [pains ASPECT] than [Motrin OBJECT].*

Data format

Представленные файлы данных имеют формат CoNLL. Каждая строка содержит одно слово и его метку, разделенные табуляцией ("Word<TAB>label"), конец предложения отмечен пустой строкой. Метки представлены в формате BIO, где каждая из меток сущности ("Объект", "Аспект", "Предикат") предваряется префиксом "B-" или "I-", указывающим начало сущности (первое слово сущности) и внутреннюю часть объекта. сущность (второе и все последующие слова). Слова, которые не являются частью сущности, помечаются буквой "O".:

advil B-Object

works O

better B-Predicate

for O

body B-Aspect

aches I-Aspect

and O

pains B-Aspect

than O

motrin B-Object

1.2 Формулировка задачи

Данные состоят из сравнительных предложений (т.е. предложений, содержащих сравнение двух или более объектов). Они содержат три типа объектов:

- **Объекты** – объекты, которые сравниваются
- **Аспекты** – характеристики, по которым сравниваются объекты
- **Сказуемое** – слова или фразы, которые реализуют сравнение (обычно сравнительные прилагательные или наречия)

В наборе данных используется схема BIO:

- Первое слово сущности помечается тегом “B-<entity-type>” (начало сущности).
- Второе и последующие слова сущности помечаются тегом “I-<entity-type>” (внутри сущности).
- Слова, которые не являются частью сущности, помечаются тегом “O” (вне сущности).

Поэтому в нашем наборе данных используются следующие метки:

- O
- B-Object
- I-Object
- B-Aspect
- I-Aspect
- B-Predicate
- I-Predicate

Ваша задача – присвоить одну из таких меток каждому токену в тестовом наборе.

1.3 Метрики оценки

Результаты будут оценены при помощи метрики F_1-score:

$$2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{tp}{tp + 0.5(fp + fn)}$$

Мы будем учитывать баллы для всех отдельных классов, кроме тега O. Для объектов, состоящих из нескольких слов, мы используем “расслабленную” метрику: если границы прогнозируемого объекта совпадают с границами эталонного объекта, мы добавляем 1 к количеству TP (количество истинных положительных примеров). Если имеется только частичное совпадение, мы добавляем число от 0 до 1, вычисляемое как длина пересечения, разделенная на полную длину объекта.

1.4 Method

Ваша задача – обучить модель разметки последовательности на предоставленном наборе данных. Вы можете использовать любой тип трансформера, подходящий для решения этой задачи (например, модели типа BERT). Мы рекомендуем вам поэкспериментировать с различными типами инициализации векторных представлений, а также настройкой дополнительных параметров.
В контексте этого задания вы решите задачу по разметке последовательностей на наборе данных сравнительных предложений, предоставленных командой курса. Вам необходимо обучить модель и представить свое решение на CodaLab.: https://codalab.lisn.upsaclay.fr/competitions/531.

1.5 Ожидаемый результат

Предполагается, что вы:

1. **Разработаете решение задачи** и предоставите воспроизводимый код в виде jupyter notebook (можно предоставить ссылку на Google Colab[1]).:
    - вы должны использовать Python 3,
    - ноутбук должен содержать код для установки всех зависимостей,
    - ноутбук должен содержать код для загрузки всех дополнительных наборов данных, которые он использует,
    - ноутбук должен воспроизвести результаты, которые вы отправляете в CodaLab:
        i. он должен сгенерировать выходной файл требуемого формата – этот результат должен быть достигнут путем запуска вашего ноутбуку ячейка за ячейкой, ничего не меняя, включая пути к файлам!
        ii. результаты должны быть близки к результатам в CodaLab.
        iii. если эта воспроизводимость не будет достигнута, ваша оценка будет понижена.
2. **Напишете отчет,** в котором описывается метод, используемый в вашем решении, как часть вашего jupyter notebook.
3. **Отправите ваше (лучшее) решение на** CodaLab:

https://codalab.lisn.upsaclay.fr/competitions/531 так, чтобы оно появилось в турнирной таблице. Никнейм участника должен присутствовать в отчете jupyter notebook для проверки. Вам необходимо использовать “ **SberSummer2023**” в качестве названия вашей команды (Настройки -> Название команды). Каждая строка содержит одно слово и его метку, разделенные табуляцией ("Word<TAB>label"), конец предложения отмечен пустой строкой. Убедитесь, что размер выходного набора данных совпадает с размером входного набора данных. Пожалуйста, отправьте файл результатов в **zip-архиве**.

[1] https://colab.research.google.com/notebooks/intro.ipynb#recent=true

Задание 2 - Text Detoxification

2.1 Введение

Глобальный доступ к Интернету позволил распространять информацию по всему миру и дал ему много новых возможностей. В то же время, наряду с преимуществами, экспоненциальный и неконтролируемый рост пользовательского контента в Интернете способствовал распространению токсичности и разжиганию ненависти.

В направлении обнаружения оскорбительной речи в настоящий момент была проделана большая работа. Важно не только обнаруживать токсичный контент, но и бороться с ними более разумными способами. В то время как некоторые социальные сети блокируют сомнительный контент, другим решение – обнаружить токсичность во вводимом тексте и предложить пользователю неоскорбительную версию его текста. Эту задачу можно считать задачей переноса стиля, где исходный стиль токсичен, а целевой стиль нейтрален / нетоксичен. Задача переноса стиля – это задача преобразования текста таким образом, чтобы его содержание и большинство свойств оставались неизменными, а один конкретный атрибут (стиль) изменялся. Этим атрибутом может быть настроение, наличие предвзятости, степень формальности и т.д. Данная задача уже рассматривалась и решалась для английского языка, тогда как методы переноса стиля текста и детоксификации текста для русского языка ранее не изучались.

2.2 Формулировка задачи

У вас есть отличный шанс стать участником конкурса автоматической детоксикации русских текстов для борьбы с ненормативной лексикой. Такой вид переноса стиля может быть использован, например, для обработки токсичного контента в социальных сетях. В то время как для английского языка в этой области была проделана большая работа, для русского языка она еще никогда не была решена.
Мы определяем задачу детоксикации как задачу переноса стиля: от токсичного стиля к нетоксичному. Необходимо переписать предложение и сохранить контекст.

Мы определяем задачу передачи стиля следующим образом. Давайте рассмотрим два корпуса $D^{\mathrm{X}} = \{x_1, x_2, \ldots, x_n\}$ и $D^{Y} = \{y_1, y_2, \ldots, y_m\}$ в двух стилях –s^{X} (токсичном) и s^{Y} (нетоксичном), соответственно. Задача состоит в том, чтобы создать модель $f_\theta : X \rightarrow Y$, где X и Y это все возможные тексты в стилях s^{X} и s^{Y} . Задача выбора оптимальных параметров θ для f состоит в максимизации вероятности $p(y' | x, s^{Y})$ переписывания предложения x в стиле s^{X} в предложение y' которое сохраняет

контент x и имеет стиль s^Y . Параметры максимизируются на корпусах D^X и D^Y и могут быть параллельными или непараллельными. Мы сосредоточимся на переносе $s^X \rightarrow s^Y$, где s^X это токсичный стиль, а s^Y – нейтральный.

2.3 Метрики оценки

Чтобы выполнить полноценную оценку модели переноса стиля, необходимо учесть следующие факторы: оценка (i) изменения стиля текста, (ii) сохранение содержания и оценка (iii) грамматичности предложения. В большинстве работ по переносу стиля используются индивидуальные метрики для оценки трех параметров. В нашем конкурсе мы используем следующие метрики:

1) Объединенная метрика J, которая сочетает в себе критерий переноса стиля, сохранение смысла, грамматичность
2) ChF: посимвольное сравнение автоматически переписанного предложения и предложения, переписанного вручную.

2.4 Методы

В контексте этого задания вы решите задачу детоксификации текста на наборе параллельных данных, предоставленном командой курса в рамках соревнования RUSSE'2022. Вам необходимо обучить модель и загрузить свое решение на CodaLab: https://codalab.lisn.upsaclay.fr/competitions/642.

Вы можете использовать любые дополнительные методы и / или модели для переноса стиля, а также предварительно обученные модели для генерации текста (GPT, T5 и т.д.). Ниже представлены базовые подходы, которые вы можете улучшить. https://github.com/skoltech-nlp/russe_detox_2022/tree/main/baselines

Duplicate – это базовый подход, который дублирует текст подаваемый на вход, другими словами не вносит никаких изменений во входное предложение. Он представляет собой нижнюю границу производительности моделей переноса стиля, т.е. помогает проверить, что модели не ухудшают исходное предложение.

Delete – этот метод устраняет токсичные слова на основе предопределенного словаря токсичных слов. Эта идея часто используется на телевидении и в других средствах массовой информации: грубые слова вычеркиваются или скрываются специальными символами (обычно звездочкой). Основным ограничением этого метода является неполнота словарного запаса: мы не можем собрать все грубые и токсичные слова. Более того, в языке могут появиться новые оскорбительные слова и фразы, которые также могут быть объединены с различными префиксами и

суффиксами. В то же время этот метод умеет достаточно хорошо сохранять содержание, за исключением случаев, когда токсичные слова содержат смысл, необходимый для понимания всего текста.

Retrieve – данный метод направлен на улучшении качества переноса стиля. Для данного токсичного предложения мы извлекаем наиболее похожий нетоксичный текст из корпуса нетоксичных образцов. В этом случае мы получаем безопасное предложение. Однако сохранение контента зависит от размера корпуса и, вероятно, будет очень низкой.

2.5 Ожидаемый результат

Пример входных и выходных данных модели представлен ниже:

Model	Sentence
Input	не дай бог моя дочь так оденется убью н[redacted]й палкой (If, God forbid, my daughter goes out dressed like this, I'll f[redacted]g kill her with a stick)
Delete	не дай бог моя дочь так оденется убью палкой (If, God forbid, my daughter goes out dressed like this, I'll kill her with a stick)
Retrieve	не бросайте угла родного одной мы лежали больнице палате в в в те девчонкой была молодой годы (don't abandon your native corner same hospital we were ward in in in those girl was young years)

В качестве ответа ожидается переписанное токсичное предложение в более нейтральном (нетоксичном стиле). Для каждого входного предложения x_i мы ожидаем соответствующее ему переписанное предложение y_i.

Пожалуйста, отправляйте свое решение в текстовом файле *results.txt*, где на каждой строке содержится только одно переписанное предложение. Убедитесь, что размер выходного набора данных совпадает с размером входного набора данных. Пожалуйста, отправьте файл результатов в **zip-архиве** в Codalab: https://codalab.lisn.upsaclay.fr/competitions/642.

Предполагается, что вы:

1. **Разработаете решение задачи** и предоставите воспроизводимый код в виде jupyter notebook (можно предоставить ссылку на Google Colab[2]).:
    - вы должны использовать Python 3,
    - ноутбук должен содержать код для установки всех зависимостей,
    - ноутбук должен содержать код для загрузки всех дополнительных наборов данных, которые он использует,

[2] https://colab.research.google.com/notebooks/intro.ipynb#recent=true

- ноутбук должен воспроизвести результаты, которые вы отправляете в CodaLab:
    i. он должен сгенерировать выходной файл требуемого формата – этот результат должен быть достигнут путем запуска вашего ноутбуку ячейка за ячейкой, ничего не меняя, включая пути к файлам!
    ii. результаты должны быть близки к результатам в CodaLab.
    iii. если эта воспроизводимость не будет достигнута, ваша оценка будет понижена.

2. **Напишете отчет,** в котором описывается метод, используемый в вашем решении, как часть вашего jupyter notebook.
3. **Отправите ваше (лучшее) решение на CodaLab**: https://codalab.lisn.upsaclay.fr/competitions/642 так, чтобы оно появилось в турнирной таблице. Никнейм участника должен присутствовать в отчете jupyter notebook для проверки. Вам необходимо использовать " **SberSummer2023**" в качестве названия вашей команды (Настройки -> Название команды).

Критерии оценки

Технический отчет		**Код**		**Результаты**		**Всего***
Методология	Анализ результатов	Читаемость	Воспроизводимость	Преодоление базового решения	top-1: 10 баллов top-20%: 5 баллов	**100% + бонус**
5	5	5	5	5 or 10	0 or 5 or 10	**25** + (5 or 10)

* Чтобы получить 100% за каждое задание, вам нужно набрать 25 баллов, но вы можете получить дополнительные 5 баллов, если ваш метод входит в топ-20% (среди всех зачисленных студентов) в таблице лидеров Codalab, и дополнительные 10 баллов, если ваш метод является топ-1 в таблице лидеров Codalab. Эти баллы будут засчитаны пропорционально итоговой оценке по курсу.

Для каждого домашнего задания вам необходимо предоставить:

1. **Технический отчет (10 баллов всего).** Напишите отчет в предоставленном шаблоне Ipynb[3] с описанием метода, используемого в вашем решении. Отчет должен состоять из двух частей:
    a. **Методология (5 баллов)**: основная часть вашего отчета с описанием всех методов, которые вы протестировали и которые сработали для вас лучше всего. Сюда вы можете включить описание препроцессинга, описание моделей и мотивации их использования, описание деталей процесса обучения (разделение на тренировочные и тестовые данные, перекрестная проверка и т.д.). Здесь должно быть все то, что поможет понять объем проделанной вами работы и воспроизвести ваш метод при отсутствии кода.
    b. **Анализ результатов (5 баллов)**: необходимо предоставить итоговую таблицу со сравнением базового решения и всех опробованных вами подходов. Даже если какой-то метод не вывел вас в топ турнирной таблицы, вы все равно должны указать этот результат и проанализировать, почему, по вашему мнению, сработал тот или иной подход. Интересные выводы в ходе рассуждений будут плюсом.

2. **Code (10 points total)**. Разработайте самостоятельно решение задачи и предоставьте воспроизводимый код в предоставленном шаблоне. Убедитесь, что ваш код:
    a. Использует Python 3;
    b. Содержит код для установки всех зависимостей;
    c. Содержит код для загрузки всех используемых наборов данных;
    d. Содержит код для воспроизведения ваших результатов (другими словами, если кто-то откроет ваш ноутбук, он должен иметь возможность запускать код по ячейкам и воспроизвести ваши результаты экспериментов).

    В результате ваш код будет оценен в соответствии со следующими критериями:
    a. **Читаемость (5 баллов)**: ваш код должен быть хорошо структурирован, предпочтительно с указанием частей вашего подхода (Предварительная обработка, обучение модели, оценка и т.д.).
    b. **Воспроизводимость (5 баллов)**: ваш код должен быть воспроизведен без каких-либо ошибок в режиме “Выполнить все” (получение экспериментальной части).

    Результаты (5 баллов + 5 или 10 баллов): Отправьте ваше (лучшее) решение на платформу **CodaLab**. Ваше имя / никнейм должны присутствовать в отчете

[3] https://colab.research.google.com/drive/1ttPT6X4K0ovgbzmNjlcEiprkj1LaBuF2#s=

для идентификации вашего решения. Вам необходимо использовать “Sber” в качестве названия вашей команды (Настройки -> Название команды).

- За преодоление базового решения вы получите **5 баллов**; далее вы можете получить дополнительные **5 баллов за позицию top-20%** в турнирной таблице на тестовых данных (private test, post-evaluation) ИЛИ **дополнительные 10 баллов за позицию top-1** на тестовых данных (private test, post-evaluation.

Дополнительные примечания:

Пожалуйста, следуйте правилам, описанным ниже:

Вы должны самостоятельно поработать над моделью и предоставить свое собственное решение.

1. Использование данных:
    a. Вы можете использовать только те размеченные данные, которые представлены в соревновании CodaLab.
    b. Вы можете использовать любые неразмеченные наборы данных, которые вам нужны, при условии, что они открыты. В отчете вы должны указать дополнительные данные, которые вы используете.
2. Для того, чтобы получить полный балл, вы должны представить свое решение до дедлайна.